Data analysis studies on the HIV infection

István Bartha

Eötvös Loránd University, Doctoral School of Biology

Head: Dr. Anna Erdei

Theoretical and evolutionary biology doctoral programme

Head: Dr. Eörs Szathmáry

Supervisor: Dr. Viktor Müller

Department of Plant Taxonomy, Ecology and Theoretical Biology

2013

1. Introduction

HIV infection is a profound public health issue, especially in Sub-Saharan Africa. During my PhD training I performed two data analysis projects which provided insights into the epidemiology and pathogenesis of the virus.

Through multiple rounds of selection and escape, host and pathogen genomes are imprinted with signatures of evolutionary changes. Due to its rapid and error-prone replication HIV exhibits fast with-in host evolution which permits for an adaptation to the individual. Using paired full genome human and viral genetic data we mapped the sites of host selective pressure on the viral proteome and revealed host genetic factors that exert such a selective pressure on the virus.

In a second project we estimated the prevalence of HIV superinfection in a large clinical database which contained HIV sequences from routine genotyping events. HIV superinfection makes recombination of different viral strains possible, and it could elevate the speed of the disease progression of an individual. Moreover, superinfection with a drug resistant HIV strain can compromise the antiretroviral therapy. We identified superinfection cases from consensus HIV *pol* sequences sampled from patients under therapy.

2. Methods

Human genome-wide genotyping and HIV-1 full-length sequencing data were available for the study. Binary variables were created for each variable amino-acid positions, for every amino acid that was present in at least 20 HIV genomes. Human SNP imputation was performed using HapMap data as reference. Associations between all SNPs and HIV-1 amino acids were tested by logistic regression under an additive genetic model. For each amino acid

that had genome-wide significant association, we searched for independently associated SNPs by iteratively conditioning on the most significant SNP.

We used sequence data from routine genotypic tests spanning the protease and partial reverse transcriptase regions in the Virolab and EuResist databases that collated data from five European countries. Superinfection was indicated when sequences of a patient failed to cluster together in phylogenetic trees constructed with selected sets of control sequences. A subset of the indicated cases was validated by re-sequencing *pol* and *env* regions from the original samples.

3. Results

3.1. Joint association analysis of the HIV and human genomes

1. We demonstrated that mapping host selective pressure on the parasite genome using genome wide association studies is feasible and biologically relevant results can be obtained without further *a priori* knowledge on the biology of the host-parasite system.

2. We identified 40 viral amino acid sites which were under selective pressure of the *HLA* genes. Of these 25 are outside of known, published epitopes.

3. We confirmed the extensive role of the *HLA* genes in the immunological response to the HIV infection.

3.2. Detecting HIV superinfection

1. We developed an analysis pipeline that allowed us to detect dual infection in large databases by leveraging modern and reliable approaches

of phylogenetic inference.

2. We identified potentially superinfected patienst in the EuResist and Virolab databases. We estimate an upper bound of 2% for the prevalence of superinfection in this population.

3. We revealed a high frequency of sample mixups which indicates the need for validation by re-sequencing in the detection of superinfection.

3.3. Subtasks in other studies

1. Estimating the life span of resting T-cells. Using bootstrap replicates I estimated the error of the estimated turnover of the cell population in case the model which is fitted on the labeling data contains a different number of compartments than the real population.

2. Identifying potential APOBEC3G induced hypermutations in HIV sequences. I found the imprints of a possible hypermutation in the RT and Gp120 proteins.

3. Creating a visualisation platform of gene expression data. The study recovered cellulare transcriptome over 24 hours after HIV infections from CD4+ T-cells. I developed a web resource which allows the querying and browsing of the expression data `http://peachi.labtelenti.org`.

4. Publications

Journal articles Bullets indicate those articles which are incorporated in the main text of the thesis.

1. **Bartha I**, Simon P, Müller V: Has HIV evolved to induce immune pathogenesis? Trends In Immunology 2008, 29:322–8.

2. • Snoeck J, Fellay J, **Bartha I**, Douek DC, Telenti A: Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology 2011, 8:87.

3. Rovó P, Stráner P, Láng A, **Bartha I**, Huszár K, Nyitray L, Perczel A: Structural insights into the Trp-cage folding intermediate formation. Chemistry 2013, 19:2628–40.

4. • Mohammadi P, Desfarges S, **Bartha I**, Joos B, Zangger N, Muñoz M, Günthard HF, Beerenwinkel N, Telenti A, Ciuffi A: 24 hours in the life of HIV-1 in a T cell line. PLoS pathogens 2013, 9:e1003161.

5. • Westera L, Drylewicz J, den Braber I, Mugwagwa T, van der Maas I, Kwast L, Volman T, van de Weg-Schrijver EHR, **Bartha I**, Spierenburg G, Gaiser K, Ackermans MT, Asquith B, de Boer RJ, Tesselaar K, Borghans J a M: Closing the gap between T-cell life span estimates from stable isotope-labeling studies in mice and men. Blood 2013.

6. • **Bartha, I.**, Carlson J., Brumme C., McLaren P., Brumme Y., John M., Haas D., Martinez-Picado J, Dalmau J., López Galíndez C., Casado C., Rauch A., Günthard H., Bernasconi E., Vernazza P., Klimkait T., Yerly S., O'Brien S., Listgarten J., Pfeifer N., Lippert C., Fusi N., Kutalik Z., Allen T., Müller V., Harrigan P., Heckerman D., Telenti A. Fellay J. . A Genome-to-Genome Analysis of Associations between Human Genetic Variation, HIV-1 Sequence Diversity, and Viral Control. Accepted for publication in eLife.

7. • **Bartha I**, Assel M., Sloot P., Zazzi M., Torti C., Schülter E., De Luca A., Sönnerborg A., Abecasis A., Van Laethem K., Rosi A., Svard J., Paredes R., van de Vijver D., Vandamme A., Müller V . Superinfection

with drug-resistant HIV is rare and does not contribute substantially to therapy failure in a large European cohort - Under review in BMC Infectious Diseases.

Oral presentations

1. 20th International HIV, Dynamics & Evolution Conference in Utrecht, Netherlands, 8-11 May 2013. Joint Association Analysis Of Genome-wide Human And Hiv-1 Variation

2. 19th Conference on Retroviruses and Opportunistic Infections (CROI 2012) in Seattle, WA USA, 5-8 March 2012. I-1003: Joint Association Analysis of Genome-Wide Human and HIV-1 Variation.

3. XVIII. AIDS 2010 International Conference in Vienna, Austria, 18-23 July 2010. Estimating the frequency of superinfection in a large European collaborative HIV database.

Posters

1. 20th International HIV, Dynamics & Evolution Conference in Utrecht, Netherlands, 8-11 May 2013. Hiv Superinfection Does Not Contribute To Transmitted Drug Resistance. **Bartha I**, Assel M., Sloot P., Zazzi M., Torti C., Schülter E., De Luca A., Sönnerborg A., Abecasis A., Van Laethem K., Rosi A., Svard J., Paredes R., van de Vijver D., Vandamme A., Müller V.

2. IEEE International Conference on Bioinformatics and Biomedicine Workshops in Atlanta, GA USA, 12 November 2011. Joint analysis of host and pathogen genomes. **I Bartha**, A Telenti, V Müller, J Fellay

3. 63rd American Society of Human Genetics Annual Meeting in Boston, MA October 22-26, 2013. Joint association analysis of genome-wide human and HIV-1 variation. **I Bartha**, J. Carlson, P.J. McLaren, Z. Brumme, Ch. Brumme, R. Harrigan, A. Rauch, H. Günthard, M. John, D. Heckerman, T.M. Allen, C.L. Galindez, J. Martinez-Picado, V. Müller, A. Telenti, J. Fellay, HIV Genome to Genome Study - Poszter elfogadva.